Making new genetic diagnoses with old data

Official journal of the American College of Medical Genetics and Genomics **ORIGINAL RESEARCH ARTICLE** Genetics inMedicine

Open

Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders

Caroline F. Wright, PhD[1,2], Jeremy F. McRae, PhD[1], Stephen Clayton, MRes[1], Giuseppe Gallone, PhD[1], Stuart Aitken, PhD[3], Tomas W. FitzGerald, PhD[1], Philip Jones, MSc[1], Elena Prigmore, PhD[1], Diana Rajan, MSc[1], Jenny Lord, PhD[1], Alejandro Sifrim, PhD[1], Rosemary Kelsell, PhD[1], Michael J. Parker, PhD[4], Jeffrey C. Barrett, PhD[1], Matthew E. Hurles, PhD[1], David R. FitzPatrick, DM[4] and Helen V. Firth, DM[1,5]; on behalf of the DDD Study[1]

Purpose: Given the rapid pace of discovery in rare disease genomics, it is likely that improvements in diagnostic yield can be made by systematically reanalyzing previously generated genomic sequence data in light of new knowledge.

Methods: We tested this hypothesis in the United Kingdom–wide Deciphering Developmental Disorders study, where in 2014 we reported a diagnostic yield of 27% through whole-exome sequencing of 1,133 children with severe developmental disorders and their parents. We reanalyzed existing data using improved variant calling methodologies, novel variant detection algorithms, updated variant annotation, evidence-based filtering strategies, and newly discovered disease-associated genes.

Results: We are now able to diagnose an additional 182 individuals, taking our overall diagnostic yield to 454/1,133 (40%), and another 43 (4%) have a finding of uncertain clinical significance. The majority of these new diagnoses are due to novel developmental disorder–associated genes discovered since our original publication.

Conclusion: This study highlights the importance of coupling large-scale research with clinical practice, and of discussing the possibility of iterative reanalysis and recontact with patients and health professionals at an early stage. We estimate that implementing parent–offspring whole-exome sequencing as a first-line diagnostic test for developmental disorders would diagnose > 50% of patients.

Genet Med advance online publication 11 January 2018

Key Words: diagnostic yield; exome sequencing; reanalysis; reclassification; recontact

INTRODUCTION

The relative affordability and accessibility of next-generation sequencing have facilitated the development of family-based genomic analysis, resulting in an explosion of gene discovery and diagnosis for rare diseases.[1–3] Diagnosis rates—here defined as the confident causal association of a genotype with the presenting phenotype—vary from 20 to 60% depending on numerous factors, including specificity of the clinical presentation, genetic heterogeneity of the disease, patient recruitment criteria, sequencing technology and analytical workflow, evidence of de novo occurrence of causal variants, and date of publication.[4–6] The latter in part reflects the accelerated rate of analytical tool development and gene discovery catalyzed by next-generation sequencing.[7] Given the pace of change throughout the field, some diagnostic variants must be presumed to be unrecognized during the initial analysis of genomic data, and without intervention, may remain undiscovered. Systematic, retrospective reanalysis of genomic data is therefore likely to improve diagnostic yield.[8] However, the logistical challenges of performing regular reanalyses, coupled with reinterpretation of the results and recontacting of clinicians and patients, are substantial.[9] To date, although several small-scale examples of this approach exist,[10,11] no large-scale diagnostic reanalyses have been published, so the potential benefits of this methodology when applied systematically across an entire cohort are currently unquantified.

Due to the extremely large number of variants in every genome, evidence-based filters are applied to prioritize potentially relevant variants for individual clinical cases. A balance must be struck between sensitivity and specificity to find potential diagnoses without being overwhelmed by false positive results. As a result, there are numerous reasons why diagnostic variants might not be recognized during the analysis of genomic data, e.g., technical failure to detect a variant in the data, incorrect annotation, limited knowledge of

[1]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, Hinxton, UK; [2]University of Exeter Medical School, Institute of Biomedical and Clinical Science, Royal Devon & Exeter Hospital, Exeter, UK; [3]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK; [4]The Wellcome Centre for Ethics and Humanities/Ethox Centre, Nuffield Department of Population Health, University of Oxford, Oxford, UK; [5]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, UK. Correspondence: Caroline F. Wright (caroline.wright@exeter.ac.uk)

Submitted 11 July 2017; accepted 20 November 2017; advance online publication 11 January 2018. doi:10.1038/gim.2017.246

Table 1 Potential analytical sources of missed diagnoses and corresponding improvements made to the DDD workflow since 2014

Step	Purpose	Potential sources of missed diagnoses	Changes to DDD workflow
Variant detection	Sequence data is mapped to the human genome reference, and variation called relative to that reference	Low-depth sequence data Incorrect reference sequence Incorrect mapping Variant detection/genotyping failed Variant class not considered (e.g., triplet repeats)	Updated versions of BWA, SAMtools, GATK, and DeNovoGear Multisample variant calling Additional variant detection algorithms
Variant annotation and filtering	Stringent filters are applied to exclude low-quality, common, and noncoding variants that are unlikely to be clinically relevant	Low-quality variant discarded Incorrect annotation of allele frequency Incorrect annotation of consequence Variant filtering thresholds too stringent	Updated version of VEP Updated MAF data Updated filtering thresholds (lower MAF, exclusion of benign inherited missense variants)
Gene prioritization	Evidence-based, disease-specific "virtual" gene panels are applied to limit variants to those with a relevant genotype (heterozygous/homozygous) and inheritance (dominant/recessive) in proven disease-causing genes	Incorrect disease mechanism Incorrect inheritance or family history Incomplete penetrance Phenotype not recorded Known gene missing from panel Causal gene not yet discovered	Updated DDG2P (November 2013 freeze used previously; June 2016 freeze used here, including 286 additional genes) Plausibly pathogenic variants shared via DECIPHER Research Track Reviewed parental phenotypes
Clinical assessment	Clinical assessment of the pathogenicity and contribution of specific variants to disease in a specific individual/family	Patient phenotype differs from previously published cases Phenotype not yet developed Evidence for pathogenicity is unclear	Candidate variants re-reviewed by core DDD clinical team and/or referring clinician Some patients clinically assessed again

BWA, Burrows–Wheeler Aligner; DDD, Deciphering Developmental Disorders study; DDG2P, Developmental Disorder Gene-to-Phenotype database; GATK, Genome Analysis Toolkit; MAF, minor allele frequency; VEP, Variant Effect Predictor.

the causative loci, or inappropriate exclusion of a variant (**Table 1**).[10] It is therefore beholden upon researchers involved in large-scale translational research studies to consider re-evaluating their protocols and reanalyzing their data, and also on clinical services to consider how reinterpreting data, reclassifying variants, and recontacting patients can best be managed.

The Deciphering Developmental Disorders (DDD) study (http://www.ddduk.org) provides an ideal cohort for developing and testing how such an iterative model of reanalysis and re-reporting might work at scale. The DDD study is a United Kingdom–wide collaboration between the National Health Service (NHS) Regional Genetics Services across the United Kingdom and Ireland and the Wellcome Trust Sanger Institute, which aims to both delineate the genetic architecture of developmental disorders and improve the diagnosis of these disorders in clinical practice using high-throughput genetic technologies. From April 2011 to 2015, the DDD study recruited ~13,500 families with severe, undiagnosed developmental disorders, including ~10,000 complete parent–offspring trios, all of whom have had all known coding genes sequenced (exome sequencing). In addition to conducting large-scale, statistical research into novel genetic causes of developmental disorders,[12,13] the DDD study also returns plausible diagnostic results to individual families via ~200 referring consultant clinical geneticists, who are responsible for their ongoing care.[14] The identification and communication of plausible diagnostic variants from the DDD study was initially designed to be conservative, to maximize positive predictive value while avoiding incorrect diagnosis, with the expectation that the methodology would be largely automated and improved iteratively throughout the study in light of new data and knowledge. An important question is therefore how much of an improvement in diagnostic yield is achievable in a clinically ascertained cohort over time. Here, we reanalyze the data from the first 1,133 family trios recruited into the study, describe improvements in the analysis and interpretation workflow, and compare the findings with our initial analysis of this cohort from 3 years earlier.[14]

MATERIALS AND METHODS

Patient recruitment and assays

Children with severe undiagnosed neurodevelopmental disorders, and/or congenital anomalies, abnormal growth parameters, dysmorphic features, and unusual behavioral phenotypes, were recruited with their parents from 24 regional genetics services across the United Kingdom and Ireland.[12,14] Specific clinical data (growth, development, family and pregnancy history, previous investigations, clinical photographs) and Human Phenotype Ontology terms[15] were recorded by the regional clinical teams for the child and parents via a secure online portal within the DECIPHER database.[16]

Saliva and/or blood-extracted DNA samples were analyzed at the Wellcome Trust Sanger Institute using whole-exome

sequencing of the family trio (Agilent SureSelect 55 MB Exome Plus with Illumina HiSeq) and exon-resolution microarray analysis of the proband (Agilent 2 × 1 M array CGH (Santa Clara, CA)).[12] A selection of candidate variants with low-quality metrics were subsequently validated using targeted Sanger sequencing.

Variant detection and annotation

Mapping of short-read sequences was carried out using the Burrows–Wheeler Aligner (version 0.59)[17] algorithm with the GRCh37 1000 Genomes Project phase 2 reference. The Genome Analysis Toolkit (GATK; version 3.1.1)[18] and SAMtools (version 0.1.19)[19] was used for sample-level BAM improvement and multisample variant calling across all samples. Ensembl Variant Effect Predictor[20] based on Ensembl gene build 76 was used to annotate variants. The population prevalence (minor allele frequency) of each variant was annotated using the Exome Aggregation Consortium (ExAC),[21] 1000 Genomes Project,[22] and internal data from all unaffected (developmentally normal) DDD parents in the cohort.

Numerous bespoke algorithms were also developed to detect specific types of additional variation: DeNovoGear [23] was used to predict likely de novo single-nucleotide variants (SNVs) and small insertions/deletions (indels) in the child, augmented with candidate de novo indels called by GATK and present in the child but not their parents; CNsolidate, CoNVex, and CIFER were used respectively to detect copy-number variants (CNVs) in the array CGH and exome data, and to predict their inheritance (unpublished data); UPDio[24] was used to detect uniparental disomy (UPD); triPOD[25] was used to detect structural mosaicism; a chromosome read-depth counter was used to detect chromosomal aneuploidy (unpublished data); and Indelible was used to detect soft-clipped reads caused by midsized indels (unpublished data). All annotated SNVs, indels, and CNVs for an individual were combined into a single Variant Call Format file.

Variant filtering

An automated variant filtering pipeline was used to narrow down the number of candidate diagnostic SNVs, indels, and CNVs (**Figure 1**),[14] using the following rules for family trios:

1. *Allele frequency*. Variants must be below a series of minor allele frequency (MAF) cut-offs, using the maximum MAF of the internal and external data combined: MAF <0.0005 (0.05%) and ExAC heterozygous allele count <5 in dominant genes; MAF <0.0005 (0.05%) and ExAC hemizygote allele count $=0$ in hemizygous genes; MAF <0.005 (0.5%) in recessive genes.
2. *Predicted consequence*. Variants must be predicted to have a functional or loss-of-function consequence within a coding gene, based on the transcript with the most severe predicted consequence (longest or canonical selected where there are multiple with the same consequence), including transcript ablation, transcript amplification, splice donor, splice acceptor, stop gained, frameshift, stop lost, start lost, in-frame insertion, in-frame deletion, and missense variants.
3. *Gene and genotype*. To target the analysis toward making a primary diagnosis, variants must overlap a Confirmed or Probable gene in our curated Developmental Disorder Gene-to-Phenotype (DDG2P) database (http://www.ebi.ac.uk/gene2phenotype),[14] and the genotypes must match the allelic requirement of the gene. A version of DDG2P from June 2016 was used in this analysis. For SNVs/indels, this includes single heterozygotes in dominant genes, homozygotes and compound heterozygotes in recessive genes, and X-chromosome hemizygotes in boys in hemizygous genes. For CNVs, this includes deletions and disruptive intragenic duplications in DDG2P genes with a loss-of-function or dominant negative mechanism, whole-gene/exon duplications in genes with an increased gene dosage mechanism, and any large (>1 MB) genic deletions/duplications. SNV/CNV compound heterozygotes were also evaluated in biallelic genes.
4. *Inheritance*. Variants in the proband must be inherited in a manner that is both consistent with the family history of disease (assuming full penetrance) and the inheritance pattern of the gene (dominant/recessive/X-linked), including de novo mutations in dominant and X-linked genes (Sanger validation required if posterior probability from DeNovoGear <0.1), inherited homozygous and compound heterozygous variants in recessive genes, inherited heterozygotes in dominant genes inherited from a developmentally affected parent, maternally inherited X-chromosome variants in boys (which are

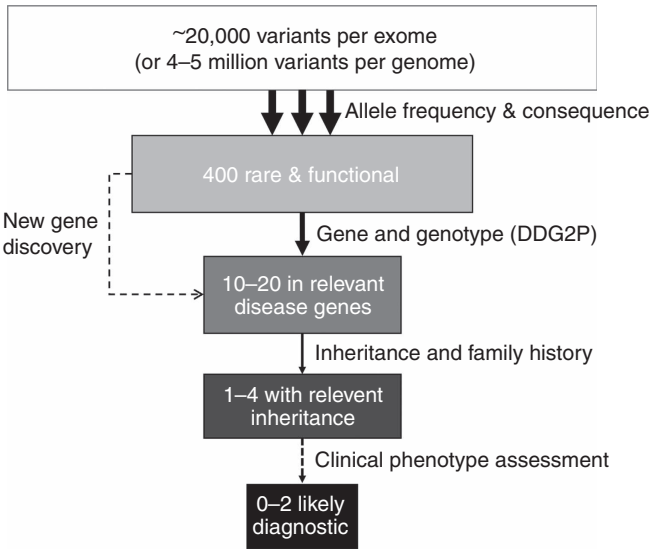

Figure 1 Outline of DDD variant filtering and reporting workflow. Details of thresholds are outlined in the Methods section. The entire workflow is automated until the final stage, which requires detailed clinical review of any candidate variants in light of the child's specific developmental phenotype. DDG2P, Developmental Disorder Gene-to-Phenotype database.

heterozygous in the mother and hemizygous in her son). Inherited missense variants predicted to be benign by PolyPhen2[26] were excluded.

Candidate variants identified through additional variant detection algorithms (including UPD, aneuploidy, structural mosaics, de novo nonessential splice sites, soft-clipped read indels, and mosaic variants inherited from unaffected parents) were analyzed and evaluated outside of this workflow.

Code availability

An updated version of the variant filtering code used by the DDD study is available online at https://github.com/jeremymcrae/clinical-filter.

Variant sharing and genetic diagnosis

Candidate diagnostic variants passing the variant filtering pipeline described above were evaluated by the DDD study's internal clinical review team (including two consultant clinical geneticists) and communicated to the regional genetics services via deposition in the DECIPHER database.[16] Both the DDD clinical team and the family's local referring NHS consultant clinical geneticist assessed the diagnostic contribution of the variant(s) to the child's presenting condition in each individual patient, based on the strength of the genetic evidence (assessment of the variant and inheritance) together with the phenotypic fit with previously reported cases. (UK NHS Consultant clinical geneticists have undertaken a minimum of 8 years training post clinical qualification including a minimum of 4 years specialist training in clinical genetics and rare disease diagnosis.) Likely diagnostic variant(s) were subsequently confirmed in an accredited diagnostic laboratory. Systematic functional studies were not performed, though all reported variants are in published developmental disorder genes with sufficient evidence to merit inclusion in our curated gene-to-phenotype database (https://www.ebi.ac.uk/gene2phenotype/).[14] Variant interpretation was informed by guidelines from both the American College of Medical Genetics and Genomics[27] and the UK Association for Clinical Genetic Science, but with the overall assessment of pathogenicity focused on an integrated clinical genetic diagnosis including a composite of patient assessment, variant evaluation, inheritance, and clinical fit. Clinical teams were asked to record the results of these evaluations in the patient's variant DECIPHER record, and anonymized variants were made publicly accessible after a short holding period.

In addition, plausibly pathogenic variants in genes not yet associated with developmental disorders, detected in children who remain undiagnosed after variant filtering, were anonymized and shared via a research track in DECIPHER, unlinked to the patient record, to facilitate variant matchmaking.[28,29] These included functional de novo variants and rare loss-of-function homozygous, compound heterozygous, and hemizygous variants in genes that are neither DDG2P nor OMIM-morbid genes. Full genomic data sets were deposited in the European Genome–Phenome Archive[30] in accordance with the Regional Ethics Committee approval for the study.

RESULTS

Using the variant detection and filtering workflow described, we have achieved a full or partial diagnosis for 454 probands in the first 1,133 family trios in the DDD study, corresponding to a 40% diagnostic yield. Of these, 78% were de novo mutations and 22% were inherited variants (12% recessively inherited from both parents, 4% dominantly inherited from an affected parent, 4% hemizygously inherited from mother to son, and 2% inherited from a mosaic unaffected parent). Thirty-three diagnoses are currently considered by the local clinical team to be a partial explanation for the child's developmental disorder (i.e., the variant explains some but not all of the child's phenotypes), while at least six probands have a dual diagnosis resulting in a compound or blended phenotype (i.e., variants in two distinct genes/loci together provide a full diagnosis for the child's condition).[11] An additional 43 probands (4%) have variants of uncertain clinical significance in known disease-associated genes, some of which may become diagnostic in future as further evidence accumulates.

The diagnostic yield increased by 13% as a result of improvements made to the workflow (**Table 1**). Overall, 182 additional probands received a new diagnosis, 272 previously diagnosed probands remained diagnosed, and 39 probands had their previous diagnoses clinically reclassified as uncertain or likely benign; a further 6 probands received a diagnosis from an independent diagnostic test that was missed by the DDD workflow due to low-depth sequencing data in at least one member of the trio. Of the new diagnoses, 35% were in 30 new disease-associated genes discovered by the DDD study itself,[12,13,31] 34% were in additional published disease genes found through literature searches, 23% resulted from improved analyses (such as updated annotations and variant filtering thresholds), and 8% resulted from additional analytical methods (**Table 2**).

A total of 838 variants were prioritized by our variant analysis and filtering workflows in this cohort, an average of ~0.7 variants per proband (**Figure 2**). Following review by two or more consultant clinical geneticists, 460 variants were classified as likely or definitely pathogenic (either fully or partially explaining the patient's phenotype, **Table 2**), versus 328 in 2014; a further 378 were classified as uncertain, likely benign, or benign for various reasons (lack of relevance of gene to phenotype, MAF too high, alternative genetic diagnosis in the proband, likely noncoding variant in the relevant transcript, analytical false positive, unrelated parental phenotype, or variant absent in affected sibling). The scale of our data set allows us to estimate the diagnostic yield of different classes of prioritized variants, which varies markedly among different inheritance modes (**Figure 3**). Over 80% of reported de novo mutations in dominant developmental disease genes, but only 10% of inherited variants in the same group of genes, were classed as likely or definitely pathogenic

Table 2 Summary of diagnoses and detection methods in the 454 diagnosed probands

Variant type	Analysis method	No. of diagnoses
Chromosomal aneuploidy	Chromosome read-depth counter	2
Copy-number variants	CNsolidate/CoNVex/CIPHER	50
De novo SNVs/indels in known genes	DeNovoGear	232
De novo SNVs/indels in new DDD genes	DeNovoGear/Discovery	58
De novo SNVs/indels in new external genes	DeNovoGear/DDD Research Variant Track	5
De novo indels in known genes	GATK candidate de novo variant	4
Inherited SNVs/indels in known genes	GATK Mendelian filter	82
Inherited SNVs/indels in new DDD genes	GATK Mendelian filter/Discovery	4
Large insertions/deletions	Soft-clipped reads	4
Mosaic structural variants	triPOD	5
Mosaic inherited SNVs/indels	Parental mosaicism	4
Nonessential splice variants	Splicing analysis	4
Uniparental disomy	UPDio	6
Total[a]	**All**	**460**

DDD, Deciphering Developmental Disorders study; GATK, Genome Analysis Toolkit; indel, insertion/deletion; SNV, single-nucleotide variant.
Reported variants that were considered by our clinical teams to explain all or part of a patient's phenotype are summarized here; the variants themselves are available with associated phenotypes through DECIPHER (https://decipher.sanger.ac.uk). All variants are in published developmental disorder genes with sufficient evidence to merit inclusion on our clinician-curated gene-to-phenotype database (https://www.ebi.ac.uk/gene2phenotype/). Note that although most variants have been analytically validated in an accredited diagnostic laboratory, functional studies have not been systematically performed to confirm clinical pathogenicity. Discovery indicates that a new developmental gene was found and published by the DDD study.[12,13,31]
[a]Includes six dual diagnoses.

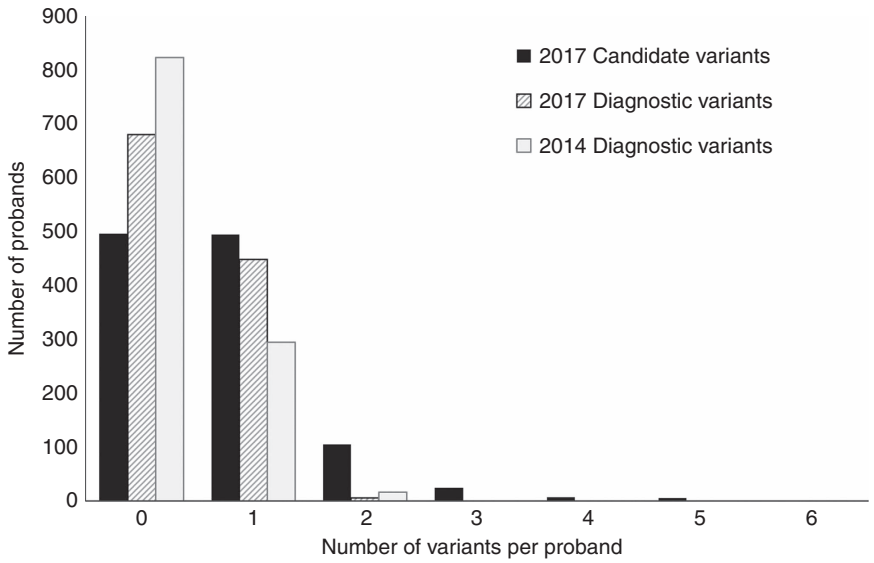

Figure 2 Summary of reported and diagnostic variants in 1,133 trios. The total number of candidate variants per proband using the 2017 analysis pipeline is indicated (black bars), along with the number of full or partially diagnostic variants per proband in 2017 (striped dark gray bars) and 2014 (light gray bars).

by our clinical teams. Of the 39 diagnoses that were reported in 2014 and have since been retracted following clinical assessment, 23 no longer meet our criteria for reporting.

The DDD study cohort excludes children who were diagnosed using standard clinical genetic testing within the NHS. Based on previous estimates of the diagnostic yield of clinical microarrays of around 10%,[32] plus a small additional diagnostic yield from single-gene testing, we estimate that the diagnostic yield of trio whole-exome sequencing would be >50% if implemented currently as a first-line test for developmental disorders.

DISCUSSION

We have developed and implemented a scalable, automated, and iterative method for reanalyzing, refiltering, re-reporting, and re-evaluating candidate diagnostic variants for severe developmental disorders from genome-wide sequence data, which in principle should be readily applicable to a wide

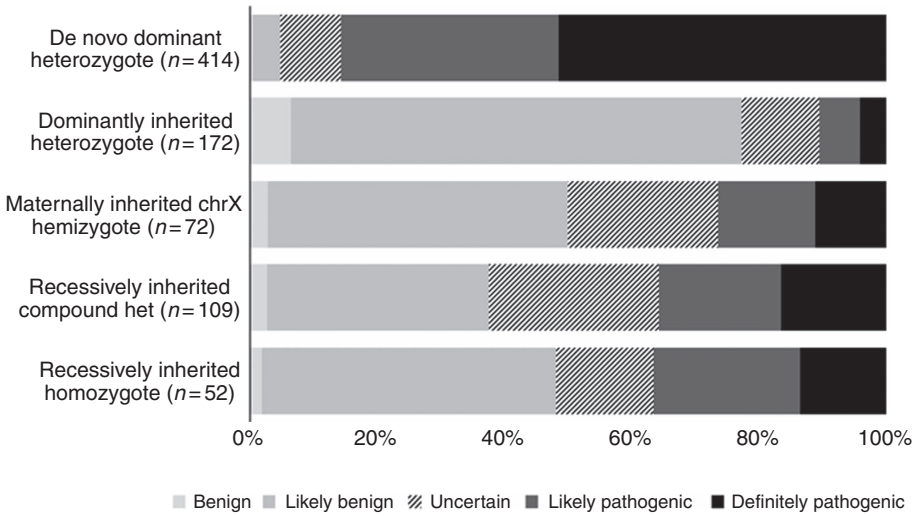

Figure 3 Pathogenicity assessments of reported variants by inheritance class. All variants (including single-nucleotide variants, indels, copy-number variants, structural variants, uniparental disomy, and aneuploidies) that were classified by clinical teams as definitely/likely pathogenic were considered diagnostic, while those considered uncertain/likely benign/benign were not. The likelihood that a rare, functional de novo mutation in a dominant DDG2P gene is considered pathogenic is >80%, while the diagnostic yield from reported inherited variants is substantially less (10–30%). Note that variants of unknown and mosaic inheritance are excluded from the diagram due to low numbers ($n < 10$).

range of rare diseases. There are numerous reasons why reassessing genomic data is necessary, and will continue to bear fruit into the future. Given the extraordinary period of rapid development and discovery in genomics, both analytical methods and variant databases become outdated very quickly. For example, considerably more background population variation data became available between our initial analysis in 2014 and this analysis in 2017 (both internally from unaffected parents within DDD, and externally from resources such as ExAC),[21] which is crucial to excluding "normal" benign variation. Furthermore, around 200–300 additional disease-causing genes are published across all rare diseases every year,[7] which are vital for finding evidence-based diagnoses within existing sequence data.

We have made a large number of evidence-based changes and upgrades to our initial variant analysis and filtering workflow within the DDD study (**Table 1**), including improved and augmented variant calling and quality control, updated variant annotation of predicted consequence and allele frequency, improved variant filtering thresholds, and additional disease-associated genes (286 additional genes were added to DDG2P between November 2013 and July 2016). Moreover, in addition to statistically well-powered gene discovery within the DDD study itself, made possible through pooling sequence data from families with developmental disorders from across the United Kingdom, we have also catalyzed gene discovery by the wider community by sharing plausibly pathogenic variants openly through the DECIPHER database. These changes have yielded substantial benefits. We are now able to diagnose an additional 182 probands in our first 1,133 trios, taking our total diagnostic yield from 27% in 2014 to 40% in 2017, highlighting the value of ongoing curation, iterative reanalysis, and re-reporting. In addition, by using an expert network of regional consultant clinical geneticists and diagnostic laboratories, we have been able to revise a small number of prior diagnoses through detailed clinical assessment. Although a variety of genetic mechanisms and inheritance patterns contribute to our diagnostic yield, ~80% of our diagnoses are de novo mutations that arose spontaneously during reproduction and are not present in either parent. Moreover, ~80% of reported de novo mutations in a known-dominant developmental disorder were classed as pathogenic by our clinical teams, emphasizing the utility of trio sequencing as a first-line strategy in sporadic cases.

Many challenges remain for continuing to improve the sensitivity and specificity of genomic sequencing. First, achieving the right balance between identifying diagnostic variants and over-reporting is problematic; the many detailed decisions required are obscured by automated workflows and hard-wired filtering thresholds. A rules-based approach will always result in reporting some false positive variants and missing some true positives. Clinical teams are usually quite unaware of which parts of the genome they are not seeing, or why, making unbiased evaluation of candidate variants extremely difficult. Moreover, variant filtering is substantially less effective for some patients and families. For family trios where both parents are unaffected and there is no family history, the majority of potentially diagnostic variants reported from exome sequencing are novel de novo mutations and are very likely to be causal; however, the converse is also true, and where both parents share a similar phenotype, the majority of reported variants are inherited and are unlikely to be causal (**Figure 3**). The situation is even more challenging

for non-trios where the parents are unavailable for testing.[14] Ever larger data sets of normal, benign variants will improve this situation, as will improved tools for predicting the pathogenicity of missense variants, but given that every family has rare/private variants, individuals and families with rare inherited dominant conditions may be better served by using more tightly focused analyses that are specific to their condition.

Second, diseases vary substantially in their genomic footprint, and those that are highly genetically heterogeneous will always be difficult to diagnose. The more genes that are causally associated with similar or overlapping phenotypes, the harder it is to be certain that any given variant is actually the cause. Although our top diagnostic genes (*ARID1B*, *SATB2*, *SCN2A*, *ANKRD11*, *MED13L*, and *SYNGAP1*) together accounted for 55 diagnoses (5% of the cohort), the substantial locus heterogeneity of developmental disorders means that most genes only contribute a single diagnosis in this cohort (**Supplementary Figure S1** online), and we have yet to find a diagnosis in the majority of the 1,400 genes on our diagnostic gene list. Although more disease-associated genes will be discovered, it is likely that these will be increasingly rare in prevalence. Substantial allelic heterogeneity also makes variant interpretation challenging even in known disease-causing genes.

Third, managing the expectations of clinicians and families is extremely challenging in such a fast-moving field, as is achieving clarity about the nature and scope of the obligations of researchers and health professionals. Diagnoses can appear at almost any time, even following a "negative report," or can be retracted as new evidence comes to light, or augmented by additional variants that may—or may not—contribute to the phenotype. Dual diagnoses resulting in blended phenotypes, which may be overlapping or distinct, are particularly challenging to untangle, as are "coincidental" findings in phenotypically heterogeneous genes where variants can cause both the disorder in question and another unrelated disorder. Although determining whether a particular variant or combination of variants explains the child's phenotype—or part of it, or none of it—is sometimes simple; other times it is not and may require further clinical evaluation and investigation. This uncertainty is the nature of a field where research and clinical practice are so entwined. By requiring peer-reviewed publication of disease-associated genes prior to addition to our diagnostic gene list and diagnostic reporting of causal variants, the DDD study has maintained a clear demarcation between research analyses and clinical practice to reduce some of this uncertainty. Through the DECIPHER platform, we also provided clinical teams with the systems and information necessary to help evaluate candidate variants. However, decisions about when and how to contact (or recontact) individual families with potential diagnoses are ultimately for local clinical teams to judge, based on their greater knowledge of the family.

Finally, a question remains as to how we should best counsel the 673 families who still have no diagnoses after several rounds of reanalyzing their data. How many more diagnoses can we expect from this same cohort in another 3 years, or another 10, and what might be reasonable for a family to expect in terms of follow-up? Large-scale sequencing studies allow us to estimate what proportion of currently undiagnosed patients are likely to be explained by a given class of variation, such as dominant de novo mutations.[13] However, in any cohort, there is likely to be a gray area between definitively genetic conditions, where a single genetic variant is the sole cause of disease, and those where multiple variants and environmental factors play a role. We don't yet know what proportion of the DDD cohort have a monogenic cause for their condition, and what fraction may have an oligogenic or polygenic component. Nonetheless, in our initial 1,133 trios, we were unable to find any statistically significant phenotypic differences between the diagnosed and undiagnosed groups (Supplementary **Figures S2** and S**3**). Currently, two-thirds of our novel diagnoses resulted from additional new disease-associated genes over the last 3 years, and it is therefore likely that the number of diagnoses will continue to increase as more causal genes are discovered through collaboration, data sharing, and meta-analyses. Although this growth in disease-associated genes is likely to slow at some point in the near future, at least for dominant diseases for which trio whole-exome study designs are very powerful, it is likely that very rare and recessive diseases will continue to be discovered for many years to come. Some diagnoses will also be missing from our data, due to low coverage in particular coding regions, long repeats or structural variants not detectable with short-read sequencing, or noncoding variants not assayed by exome sequencing. Although this suggests that whole-genome sequencing should increase our diagnostic yield further, the additional yield from genome sequencing is unlikely to be substantial given that we know of just six "missed" diagnoses in our cohort. The emphasis for future reanalysis and diagnostic reporting ought therefore to focus on better curation of gene–disease relationships and the continued coupling of research and clinical practice to enable robust gene discovery.

This work has significant implications for diagnostic laboratory reports. We suggest that iterative reinterpretation of already reported clinical sequencing data should become routine. This would require a major cultural change in reporting that would have implications for the development of appropriate informatics systems, the prioritization of clinical expertise, and the emotional burden on affected individuals and their families, all of whom may have to deal with the uncertainty of diagnoses emerging subsequently even following an initial negative report. Further work is needed to investigate the logistical and communication challenges, resource implications, and informatics infrastructure required to implement systematic reinterpretation and recontact in clinical practice.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (grant WT098051). H.V.F. is supported by the Wellcome Trust (award 200990/Z/16/Z;) "Designing, developing and delivering integrated foundations for genomic medicine"). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. We thank all the patients and families in the DDD study for their patience. Informed consent was obtained from all subjects. We also thank all the staff at the regional genetics services in the United Kingdom and Ireland, the core sample processing and analysis pipelines at the Wellcome Trust Sanger Institute, and the DECIPHER team. A full list of referring clinicians for the 1,133 families can be found in *Nature* 2015;519:223–228. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). This study uses DECIPHER (https://decipher.sanger.ac.uk), which is funded by the Wellcome Trust.

DISCLOSURE

M.E.H. is Scientific Director of Congenica. J.C.B. is Director of Open Targets. The other authors declare no conflict of interest.

REFERENCES

1. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 2013;14:681–691.
2. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–755.
3. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med* 2013;369:1502–1511.
4. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* 2015;47:717–726.
5. McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome Med* 2014;6:26.
6. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genet Med* 2016;18:696–704.
7. Chong JX, Buckingham KJ, Jhangiani SN, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 2015;97:199–215.
8. Smith ED, Radtke K, Rossi M, et al. Classification of genes: standardized clinical validity assessment of gene-disease associations aids diagnostic exome analysis and reclassifications. *Hum Mutat* 2017;38:600–608.
9. Otten E, Plantinga M, Birnie E, et al. Is there a duty to recontact in light of new genetic technologies? A systematic review of the literature. *Genet Med* 2015;17:668–678.
10. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med* 2017;19:209–214.
11. Eldomery MK, Coban-Akdemir Z, Harel T, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med* 2017;9:26.
12. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015;519:223–228.
13. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017;542:433–438.
14. Wright CF, Fitzgerald TW, Jones WD, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* 2015;385:1305–1314.
15. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* 2014;42:D966–74.
16. Bragin E, Chatzimichali EA, Wright CF, et al. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* 2014;42:D993–D1000.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
18. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
19. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
20. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17:122.
21. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–291.
22. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.
23. Ramu A, Noordam MJ, Schwartz RS, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* 2013;10:985–987.
24. King DA, Fitzgerald TW, Miller R, et al. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. *Genome Res* 2014;24:673–687.
25. King DA, Jones WD, Crow YJ, et al. Mosaic structural variation in children with developmental disorders. *Hum Mol Genet* 2015;24:2733–2745.
26. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
27. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–424.
28. Chatzimichali EA, Brent S, Hutton B, et al. Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Hum Mutat* 2015;36:941–949.
29. Wright CF, Hurles ME, Firth HV. Principle of proportionality in genomic data sharing. *Nat Rev Genet* 2016;17:1–2.
30. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-Phenome Archive of human data consented for biomedical research. *Nat Genet* 2015;47:692–695.
31. Akawi N, McRae J, Ansari M, et al. Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat Genet* 2015;47:1363–1369.
32. Sagoo GS, Butterworth AS, Sanderson S, Shaw-Smith C, Higgins JPT, Burton H. Array CGH in patients with learning disability (mental retardation) and congenital anomalies: updated systematic review and meta-analysis of 19 studies and 13,926 subjects. *Genet Med* 2009;11:139–146.